

# Sequential methods for ocean data assimilation

## From theory to practical implementations (I)

P. BRASSEUR

CNRS/LEGI, Grenoble, France  
[Pierre.Brasseur@hmg.inpg.fr](mailto:Pierre.Brasseur@hmg.inpg.fr)



J. Ballabrera, L. Berline, F. Birol, J.M. Brankart, G. Broquet, V. Carmillet, F. Castruccio, F. Debost, D. Rozier, Y. Ourmières, T.Penduff, J. Verron



Th. Delcroix, B. Dewitte, Y. duPenhoat, F. Durand, L. Gourdeau



E. Blayo, S. Carme, I. Hoteit, Pham D.T., C. Robert



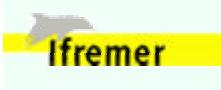
A. Barth, C. Raick



C.E. Testut, B. Tranchant



L. Parent



# OUTLINE

- **State-of-the-art**
  1. Kalman filter: fundamentals
  2. Ocean data assimilation: specific issues
  3. Error sub-spaces
  4. Low rank filters: SEEK and EnKF
- **Advanced issues**
  5. Objective validation and evaluation of DA systems
  6. Error tuning and adaptive schemes
  7. Improved temporal strategies : FGAT and IAU
  8. Kalman filtering with inequality constraints
- **The MERCATOR Assimilation System**

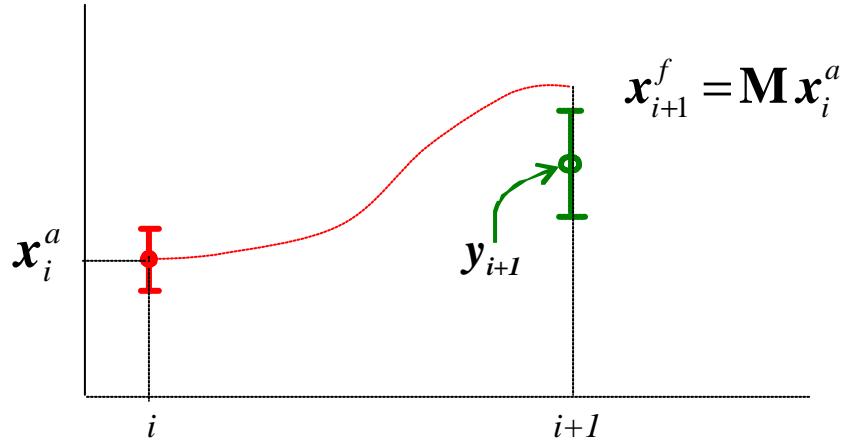
# Ocean data assimilation

- Data assimilation involves the optimal combination of measurements with the underlying dynamical principles governing the system under observation.
- Data assimilation can serve several oceanographic objectives:
  - Ocean state estimation in space & time (4D) ;
  - Detection of model errors ;
  - Estimation of budgets & model parameters ;
  - Initialisation, prediction, monitoring ;
  - Optimal design of complex observation systems ;
  - ...
- Theories: optimal control (VAR) and optimal estimation (Kalman)

## 1. Kalman Filter fundamentals

### *Problem definition*

Notations: Ide et al. (1997)



$\boldsymbol{x}_i^a$  : estimation of the « true » state vector  $\boldsymbol{x}_i^t$  at time  $i$ , dimension  $n$

$\boldsymbol{x}_{i+1}^f$  : forecast of the state vector at time  $i+1$ , using the linear model  $\mathbf{M}$

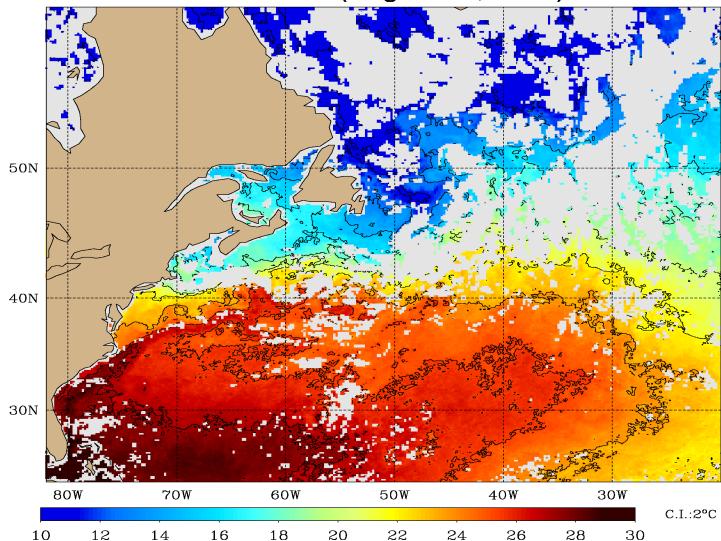
$\boldsymbol{y}_{i+1}$  : observations available at time  $i+1$ , dimension  $p$

Misfit between forecast and observations should be small :  
How « small » ?

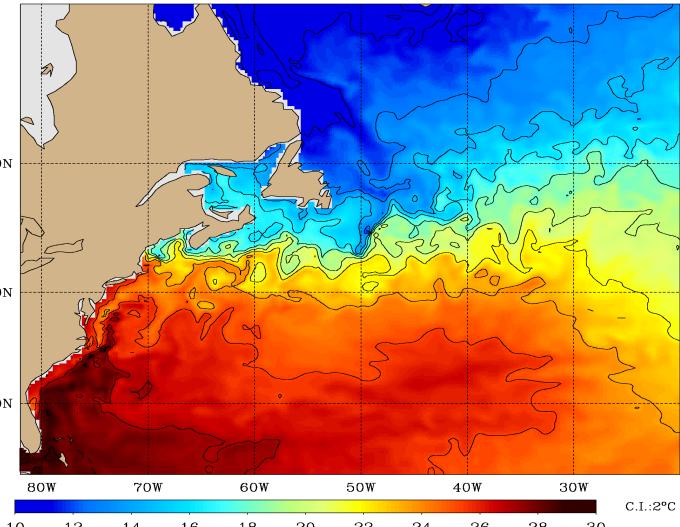
# 1. Kalman Filter fundamentals

## *Model-data misfit*

Sea Surface Temperature on Gulf Stream  
avhrr SST (August 26, 1993)



Sea Surface Temperature on Gulf Stream  
NATL3 free simulation (August 26, 1993)



$$\mathbf{y}_{i+1}$$

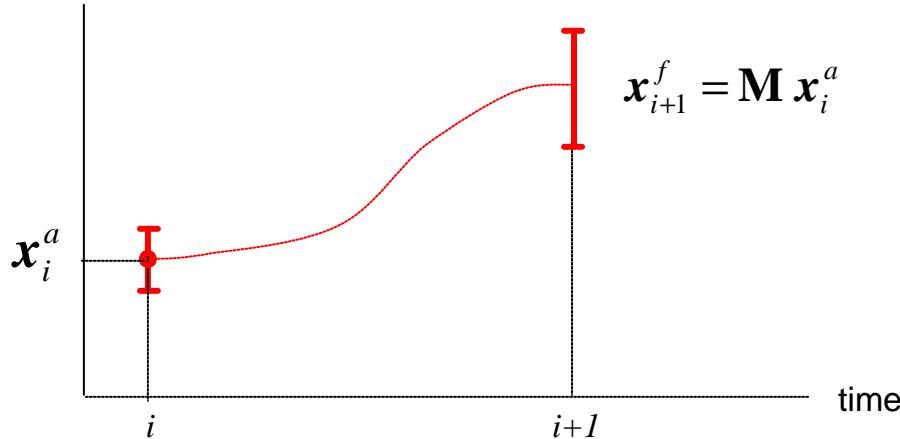
*incomplete data*

$$\mathbf{x}_{i+1}^f = \mathbf{M} \mathbf{x}_i^a$$

*imperfect model*

## 1. Kalman Filter fundamentals

### *Uncertainties & PDFs*



$e_i^a = x_i^a - x_i^t$  : error on state estimate at time  $i$  ; unknown quantity, but

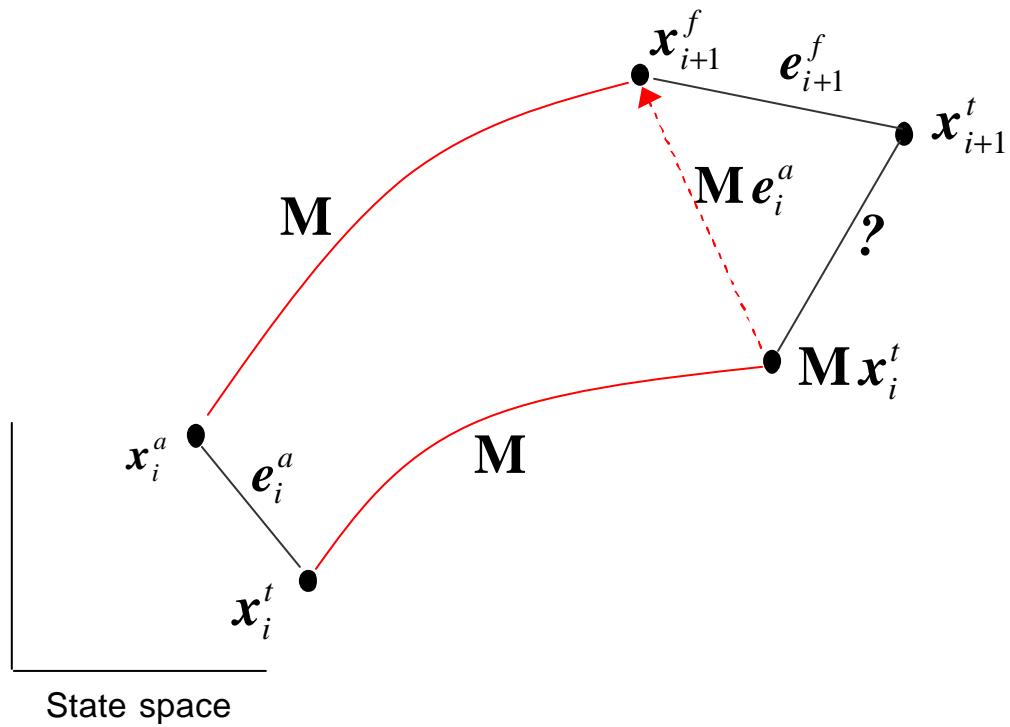
$$e_i^a \xrightarrow{\text{assume}} N(0, \mathbf{P}_i^a) \sim \exp\left[-\frac{1}{2} e_i^{aT} \mathbf{P}_i^{a-1} e_i^a\right] \quad (1)$$

? =  $\mathbf{M} x_i^t - x_{i+1}^t$  : error on model forecast between time  $i$  and time  $i+1$  ; assume:

$$? \rightarrow N(0, \mathbf{Q}) \sim \exp\left[-\frac{1}{2} ?^T \mathbf{Q}^{-1} ?\right] \quad (2)$$

## 1. Kalman Filter fundamentals

### Error diagram



## 1. Kalman Filter fundamentals

### Forecast error

$$\mathbf{M}e_i^a = \mathbf{M}x_i^a - \mathbf{M}x_i^t = x_{i+1}^f - (x_{i+1}^t + ?) = e_{i+1}^f - ?$$

Assuming pdf (1) and (2), model linearity and uncorrelated initial and modelling errors, the forecast error is distributed as :

$$e_{i+1}^f \rightarrow N(0, \mathbf{P}_{i+1}^f) \sim \exp\left[-\frac{1}{2} e_{i+1}^{f T} \mathbf{P}_{i+1}^{f^{-1}} e_{i+1}^f\right] \quad (3)$$

with

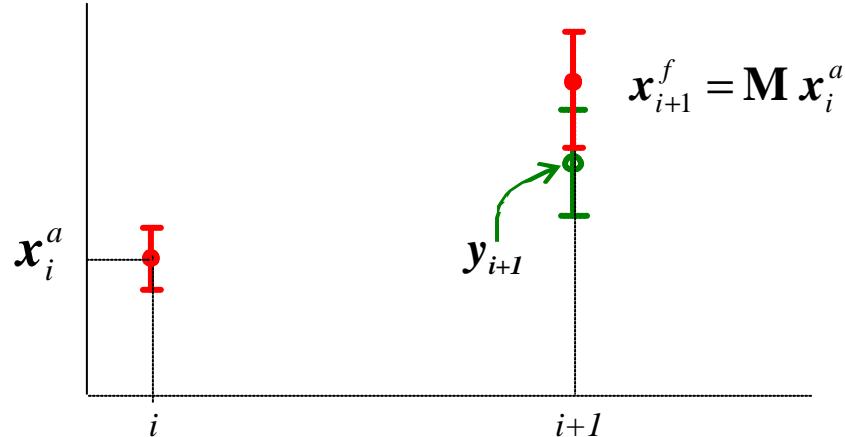
$$\begin{aligned} e_{i+1}^f &= \mathbf{M}e_i^a + ? \Rightarrow \overline{e_{i+1}^{f T} e_{i+1}^{f T}} = \mathbf{M}e_i^a e_i^{a T} \mathbf{M}^T + \overline{? ?^T} \\ &\Rightarrow \mathbf{P}_{i+1}^f = \mathbf{M} \mathbf{P}_i^a \mathbf{M}^T + \mathbf{Q} \end{aligned} \quad (4)$$

The estimation error is amplified by:

- unstable model dynamics ( $\mathbf{M}$ ) ;
- modelling errors  $\mathbf{Q}$  .

## 1. Kalman Filter fundamentals

### *Observations and errors*



$$\mathbf{y}_{i+1} = \mathbf{H} \mathbf{x}_{i+1}^t + \mathbf{e}_{i+1}^o : \text{observation available at time } i+1$$

A probability distribution for the observation error is assumed:

$$\mathbf{e}_{i+1}^o \rightarrow N(0, \mathbf{R}) \sim \exp\left[-\frac{1}{2} \mathbf{e}_{i+1}^{o \ T} \mathbf{R}^{-1} \mathbf{e}_{i+1}^o\right] \quad (5)$$

## 1. Kalman Filter fundamentals

### *Optimal estimation*

Using Bayes rule at time  $i+1$  :

$$P(\mathbf{x}_{i+1}^t | \mathbf{y}_{i+1}) = \frac{\underbrace{P(\mathbf{y}_{i+1} | \mathbf{x}_{i+1}^t)}_{\text{given by (5)}} \cdot \underbrace{P(\mathbf{x}_{i+1}^t)}_{\text{given by (3)}}}{\underbrace{P(\mathbf{y}_{i+1})}_{\text{a factor independent of } \mathbf{x}_{i+1}^t}} \quad (6)$$

$$P(\mathbf{y}_{i+1} | \mathbf{x}_{i+1}^t) \cdot P(\mathbf{x}_{i+1}^t) \sim$$

$$\begin{aligned} & \exp \left[ -\frac{1}{2} (\mathbf{x}_{i+1}^f - \mathbf{x}_{i+1}^t)^T \mathbf{P}_{i+1}^{f^{-1}} (\mathbf{x}_{i+1}^f - \mathbf{x}_{i+1}^t) \right] \cdot \exp \left[ -\frac{1}{2} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}_{i+1}^t)^T \mathbf{R}^{-1} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}_{i+1}^t) \right] \\ &= \exp \left[ -\frac{1}{2} \left\{ (\mathbf{x}_{i+1}^f - \mathbf{x}_{i+1}^t)^T \mathbf{P}_{i+1}^{f^{-1}} (\mathbf{x}_{i+1}^f - \mathbf{x}_{i+1}^t) + (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}_{i+1}^t)^T \mathbf{R}^{-1} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}_{i+1}^t) \right\} \right] \quad (7) \end{aligned}$$

The best estimate of  $\mathbf{x}_{i+1}^t$  is the value of  $\mathbf{x}$  which maximize (7), i.e. the minimum of :

$$J(\mathbf{x}) = (\mathbf{x}_{i+1}^f - \mathbf{x})^T \mathbf{P}_{i+1}^{f^{-1}} (\mathbf{x}_{i+1}^f - \mathbf{x}) + (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}) \quad (8)$$

## 1. Kalman Filter fundamentals

### *Kalman gain*

$$\mathbf{d}_x J(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{x} = \mathbf{x}_{i+1}^f + \mathbf{P}_{i+1}^f \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}) \quad (9)$$

Equation (9) can be solved for  $\mathbf{x}$  using simple algebra (\*), leading to:

$$\mathbf{x} = \mathbf{x}_{i+1}^f + \underbrace{\mathbf{P}_{i+1}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_{i+1}^f \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{y}_{i+1} - \mathbf{H} \mathbf{x}_{i+1}^f)}_{\text{Kalman gain} = \mathbf{K}_{i+1}}$$

Note: the forecast and analysis equations can be extended to *weakly* non-linear models  $\mathbf{M}$  and observation operator  $\mathbf{H}$ .

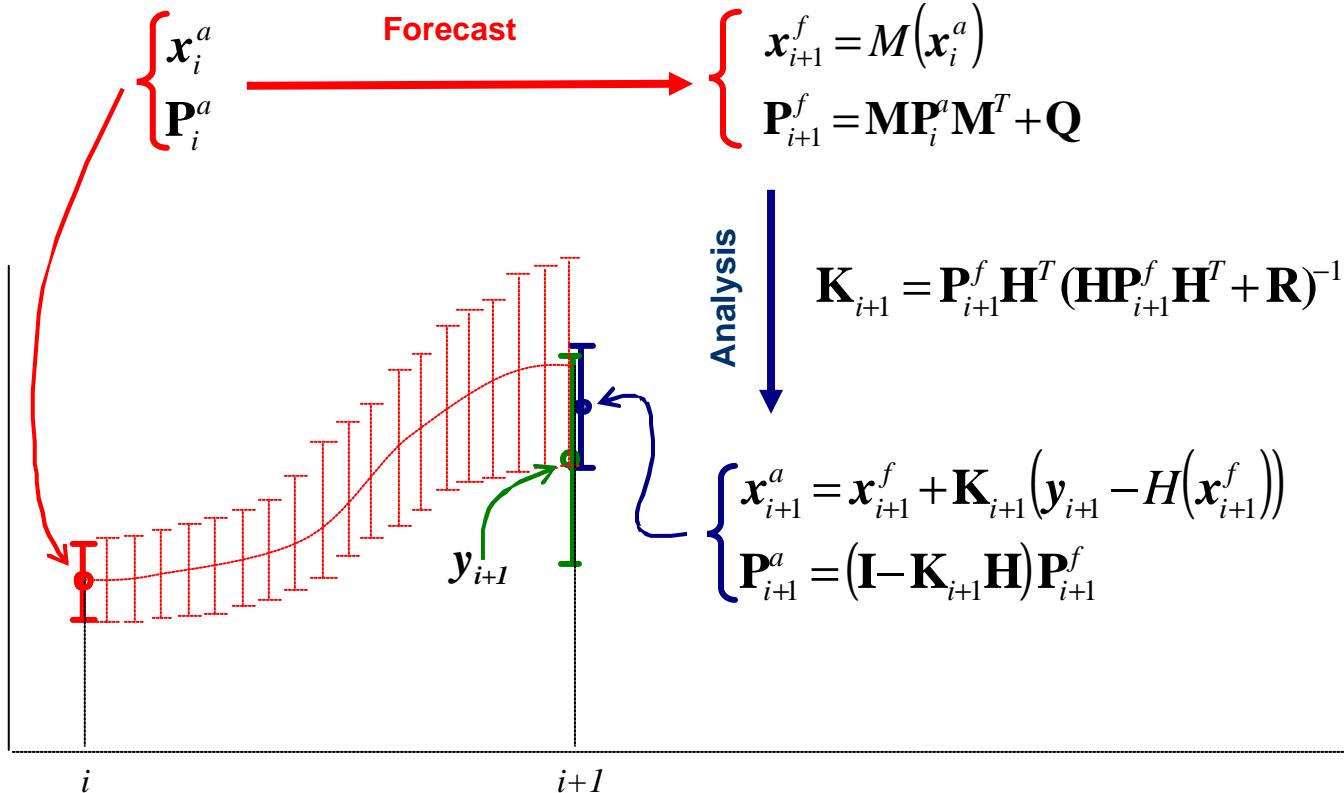
---

(\*) Hint : use matrix equality  $[\mathbf{X}_1 + \mathbf{X}_{12} \mathbf{X}_2^{-1} \mathbf{X}_{21}]^{-1} = \mathbf{X}_1^{-1} - \mathbf{X}_1^{-1} \mathbf{X}_{12} [\mathbf{X}_2 + \mathbf{X}_{21} \mathbf{X}_1^{-1} \mathbf{X}_{12}]^{-1} \mathbf{X}_{21} \mathbf{X}_1^{-1}$

---

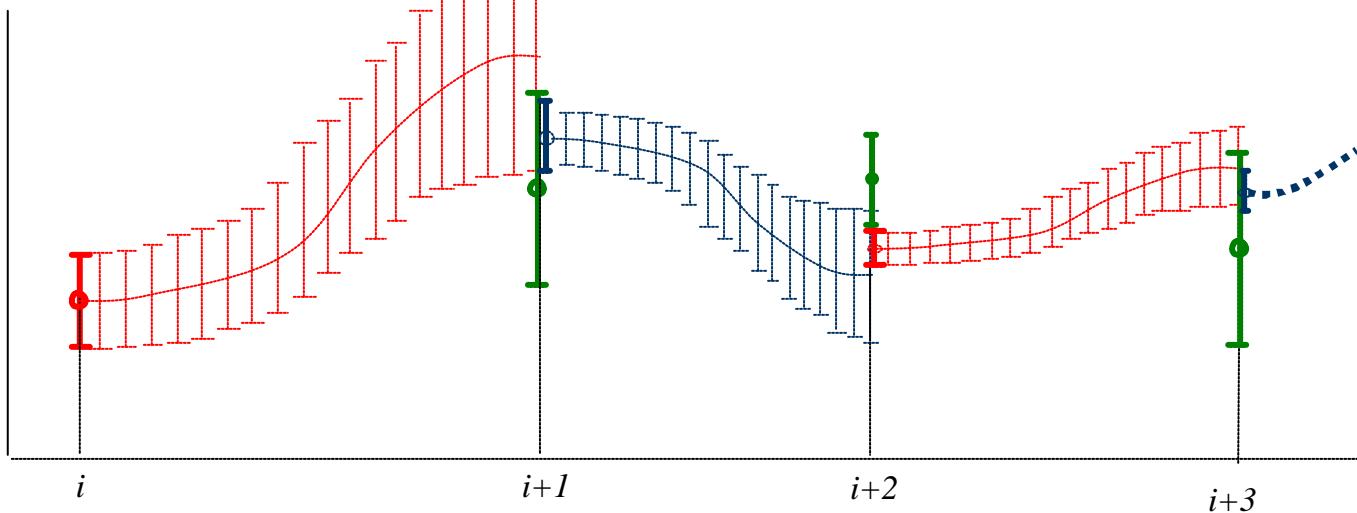
## 1. Kalman Filter fundamentals

### Assimilation cycle



## 1. Kalman Filter fundamentals Assimilation sequence

**Sequential assimilation = repeated forecast/analysis cycles**



The best estimate at a given time is influenced by all previous observations (Kalman « filter »), and the analysis error covariance reflects the competition between this accumulation of past information and the error growth due to model imperfections .

## 1. Kalman Filter fundamentals

### *Optimal Interpolation*

Noting that :

- The forecast error requires  $2n$  model integrations !!!

$$\mathbf{P}_{i+1}^f = \mathbf{M}\mathbf{P}_i^a\mathbf{M}^T + \mathbf{Q} = \mathbf{M}(\mathbf{M}\mathbf{P}_i^a)^T + \mathbf{Q}$$

- The  $2n$  model integrations are useless if model errors  $\mathbf{Q}$  are poorly known (the KF is optimal only when error statistics are perfectly known),

#### ? Simplification of the Kalman filter: « Optimal Interpolation »

To save cost and memory requirements, the KF can be simplified, using time-independent « background » covariance matrix

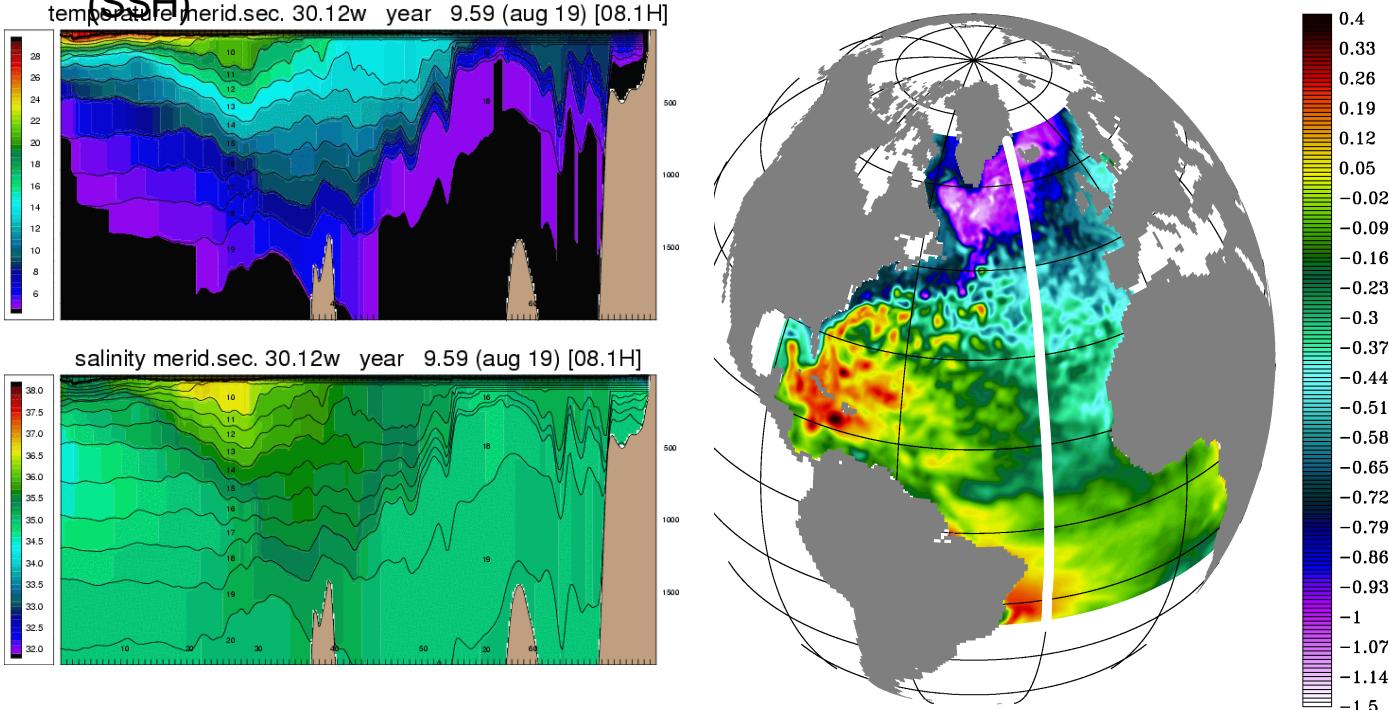
$$\mathbf{P}_{i+1}^f = \mathbf{B} = \mathbf{D}^{1/2} \underbrace{\mathbf{C}\mathbf{D}^{1/2}}_{\substack{\text{correlations} \\ \text{variances}}}$$

## 2. Ocean data assimilation : specific issues

### *Model complexity*

#### Model variables in HYCOM(\*)

temperature, salinity, velocity, layer thicknesses, sea-surface height  
(SSH)



(\*) ocean circulation model developed at Univ. Miami (RSMAS, E. Chassignet)

## 2. Ocean data assimilation : specific issues

### *State vector dimension*

- **HYCOM state vector  $x$**  : 3D grid of the 5 scalar model variables  
+ 2D grid for SSH
- **R&D prototype:**  $1/3^\circ$  horizontal resolution , 19 hybrid layers  
 $n \sim 5 \times 350 \times 350 \times 19 \sim 1.1 \times 10^7$   
**Operational prototype:**  $1/12^\circ$  horizontal resolution , 26 hybrid layers  
 $n \sim 5 \times 1400 \times 1400 \times 26 \sim 2.5 \times 10^8$
- **M operator** : dim  $n \times n \sim 6 \times 10^{16}$  real\*8 (i.e. ~ 6000 Earth Simulators !)

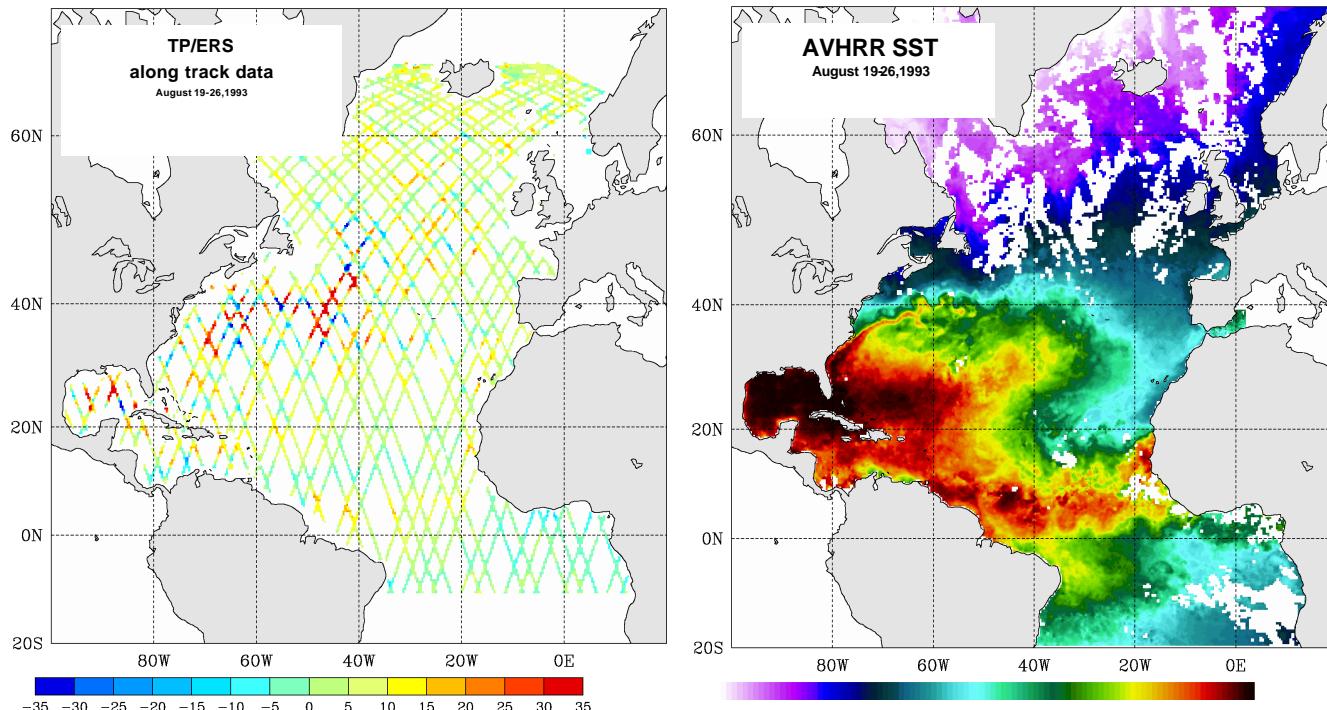
➤ The state vector dimension can be huge  
➤ The model transition matrix «  $M$  » cannot be represented explicitly.  
➤ Instead, a computer code is used to transition «  $x$  » from time  $i$  to  $i+1$

## 2. Ocean data assimilation : specific issues

### Space observations

#### Observed variables:

- from space: sea-surface height (SSH), sea-surface temperature (SST)



## 2. Ocean data assimilation : specific issues

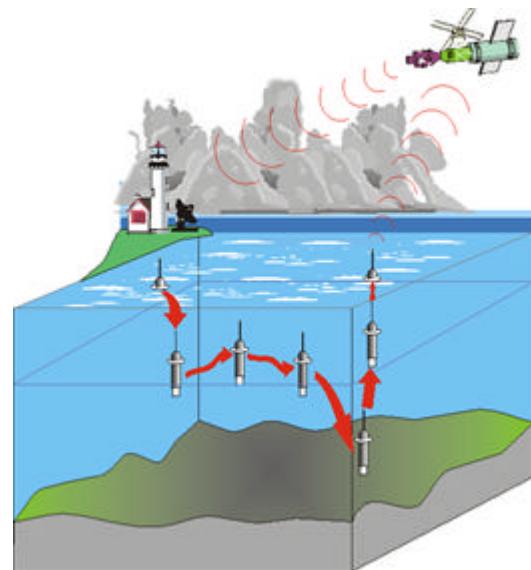
### *In situ observations*

#### Observed variables:

- *in situ*: T/S profiles (drifting floats, field campaigns, ...)



AUSTRALIA	GERMANY	NEW ZEALAND
CANADA	INDIA	NORWAY
CHINA	IRELAND	RUSSIAN FEDERATION
DENMARK	JAPAN	SPAIN
EUROPEAN UNION	KOREA (Rep. of)	UNITED KINGDOM
FRANCE	MAURITIUS	UNITED STATES



ARGO, July 2004

## 2. Ocean data assimilation : specific issues

### Observations

- **Observation vector  $y$**  : data from various sources, at different time and space resolution
  - **Radar Altimetry** : along-track measurements of SSH anomalies  
(JASON: 1 obs. / 7 km, ~ 300 km Equatorial tracks separation, repeated every 10 days ;  
ERS/ENVISAT: ~ 80 km Equatorial tracks separation, repeated every 35 days)
  - **AVHRR SST** : weekly composite images at 4 km resolution (if no clouds)
  - **ARGO floats** :  $3^\circ \times 3^\circ$  horizontal resolution (targetted), profiles (between 2000 m depth to surface) every 10 days with 1 obs / m along vertical
- $p = \dim y$  is much smaller than  $n = \dim x$  : too few observations !
  - The ocean surface relatively well observed by satellites: vertical extrapolation of data assimilated at the surface into the ocean's interior has to be consistent with vertical data profiles
  - The observed variables are closely related to model variables:  
 $\mathbf{H}$  is mainly an interpolation operator (~ simple)

## 2. Ocean data assimilation : specific issues

### *Error covariance matrix*

---

- Specification of error covariance matrix  $\mathbf{P}_0^a$  ?
- Assume a background state  $\mathbf{x}_0$  and associated error covariance  $\mathbf{P}_0$   
Consider the analysis step with only one data  $\mathbf{y}$  at a model grid point and the associated observation error  $\mathbf{e}$ .
  - $p = 1$ ,  $\mathbf{y}$  is a scalar and  $\mathbf{H}$  is a vector of the form  $\mathbf{H} = [0, \dots, 0, 1, 0, \dots, 0]$
  - The Kalman gain is then a  $(n \times 1)$  vector:

$$\mathbf{K} = \mathbf{P}_0 \mathbf{H}^T (\mathbf{H} \mathbf{P}_0 \mathbf{H}^T + \mathbf{R})^{-1} = \frac{1}{(p_{hh} + \mathbf{e}^2)} \{\mathbf{P}_0\}_h \quad \text{with} \quad p_{hh} = \{\mathbf{P}_0\}_{hh}$$

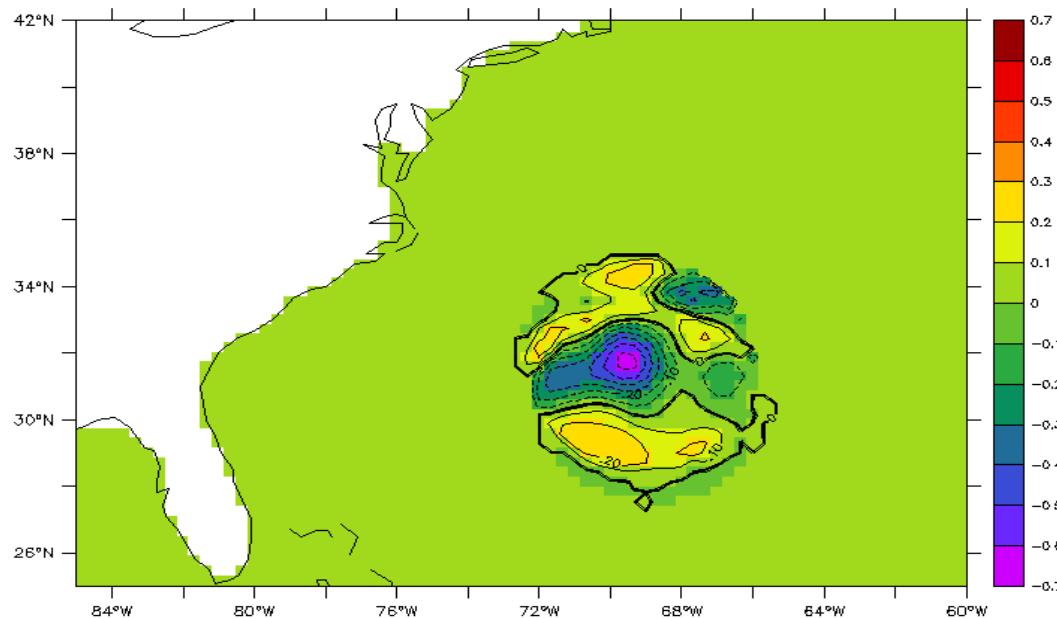
➤ The posterior estimate is a correction of the background using the  $\mathbf{h}$ -column of  $\mathbf{P}_0$

$$\hat{\mathbf{x}} = \mathbf{x}_0 + \frac{1}{(p_{hh} + \mathbf{e}^2)} \{\mathbf{P}_0\}_h (\mathbf{h} - \mathbf{h}_0) \quad \text{with} \quad \mathbf{h}_0 = \{\mathbf{x}_0\}_h$$

## 2. Ocean data assimilation : specific issues

### Horizontal covariance structures

Example: Horizontal covariance relative to a SSH ( $\eta$ ) point at (32°N,70°W)  
MERCATOR Assimilation System - Testut *et al.*(2004)

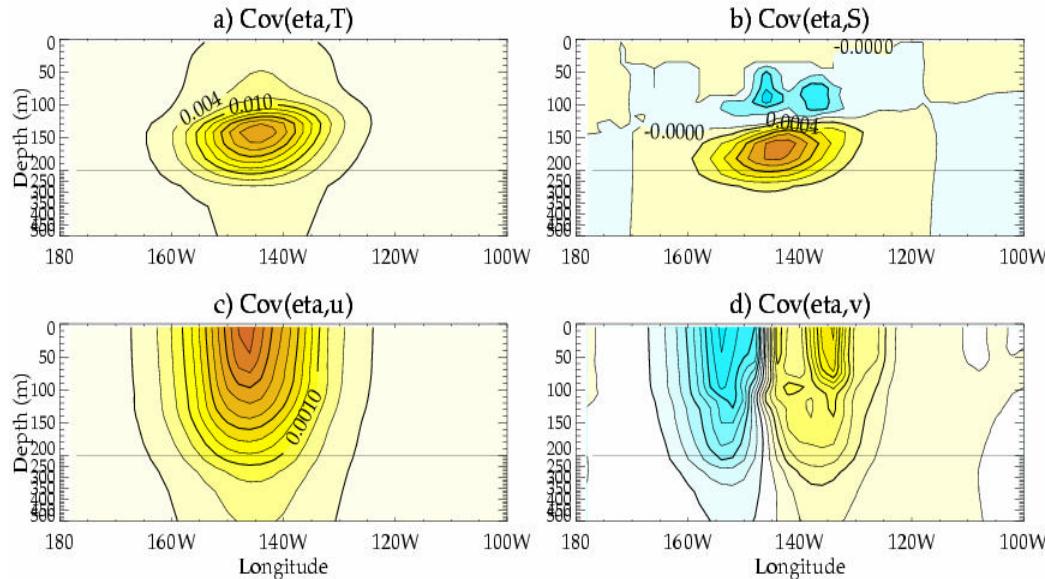


*Influence function of a single altimeter  
measurement in the sub-tropical gyre*

## 2. Ocean data assimilation : specific issues

### Multivariate covariance structures

Example: covariance relative to a SSH ( $\eta$ ) point at  $(0^\circ, 144^\circ\text{W})$  - Weaver et



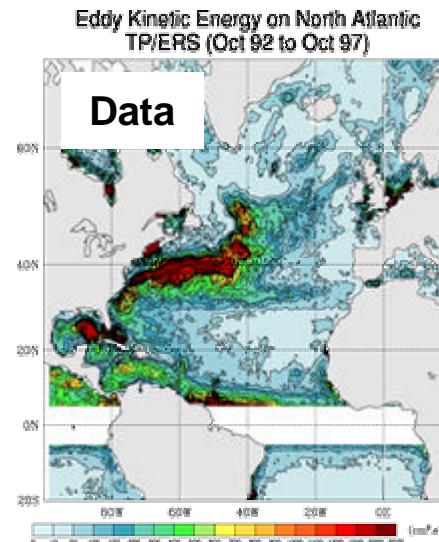
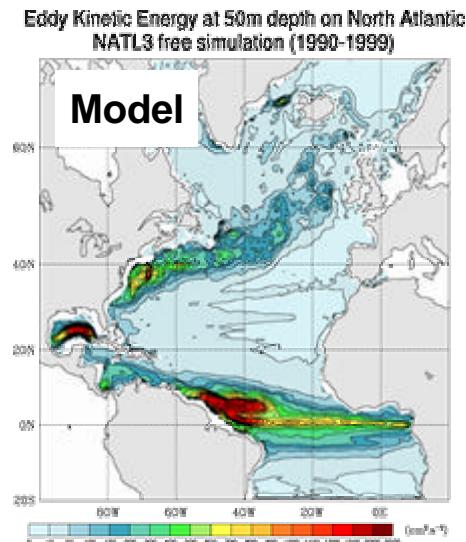
- The rows/columns of  $\mathbf{P}$  should be « balanced » dynamically.
- This requires multivariate covariances
- A full-rank representation of  $\mathbf{P}$  (dim  $n \times n$ ) is still impossible !

## 2. Ocean data assimilation : specific issues

### **Model errors Q**

#### **Model variability differs from observed variability**

EKE  
example :



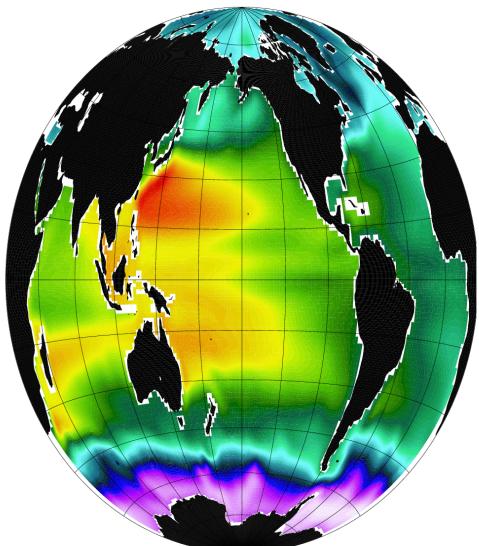
- Many different model error sources (finite discretizations, representation of bottom topography, atmospheric forcings, etc ...) which cannot be easily quantified in terms of a **Q** matrix

## 2. Ocean data assimilation : specific issues

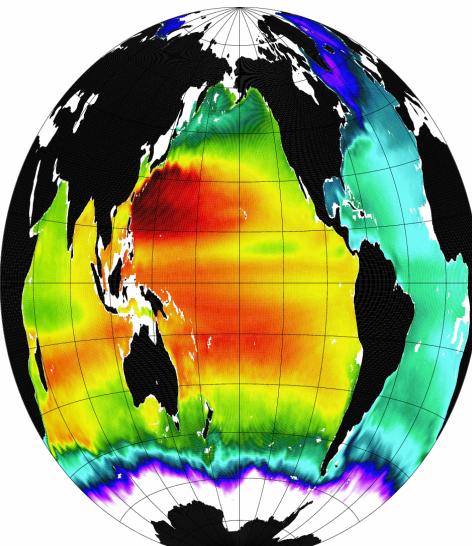
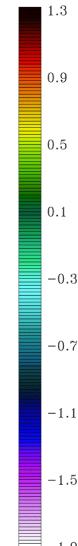
### Systematic model errors

- **Model mean SSH differs from observed mean SSH**

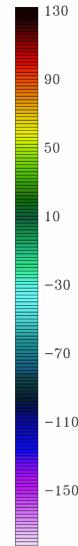
Mean sea-level difference between Pacific and Atlantic systematically too small in the model



OPA model (Madec et al. 2003)



Data (Niiler et al., 2003)



- Model biases cannot be properly taken into account with Kalman filters